A Workflow For On The Fly Normalisation Of 17th c. French

Simon Gabay, Marine Riguet, Loïc Barrault

A Workflow For On The Fly Normalisation Of 17th c. French

Simon Gabay (simon.gabay@unine.ch), Université de Neuchâtel, Switzerland

Marine Riguet (marineriguet@gmail.com), Université Paris-Sorbonne/Labex Obvil

Loïc Barrault (loic.barrault@univ-lemans.fr), Le Mans Université

Normalisation can be produced with various solutions (Baron and Rayson, 2008; Porta et al., 2013; Scherrer and Erjavec, 2013; Bollmann and Søgaard, 2016; Ljubešic et al., 2016; Tjong Kim Sang et al., 2017; Domingo et al., 2017), but recent research have demonstrated that neural machine translation (NMT) is the most efficient (Korchagina, 2017; Domingo and Casacuberta, 2018a). However, moving from test to production of a working tool is not an easy task, because of the amount of training data required for machine learning. This paper present a solution to create a parallel corpus and deliver an NMT-based normaliser for modern French.

A first test corpus

A first test has been made with the 1668 edition of *Andromaque* of Jean Racine (Racine, 1668) and the 1624 edition of the *Lettres* of Jean-Louis Guez de Balzac (Guez de Balzac, 1624).

	Author	Text	Date	Lines	Tokens	Characters
Corpus	Guez de Balzac	*Correspondance*	1624	1723	49,589	298,486
	Racine	*Andromaque*	1664	1756	13,884	86,612
Total				3479	63,473	385,098

This proto-corpus is deliberately heterogeneous to test our workflow. Guez's *Correspondance* is a collection of short letters in prose using a graphic system from the first half of the 17th c. Racine's *Andromaque* is a play in verse with a graphic system from the second half of the 17th c.

Transcriptions have been produced directly from PDF files (Fig. 1) with a model specifically designed for 17th c. prints (Gabay, 2019). It has been trained on both low-quality (72 DPI) and high-quality (400 DPI) images of books using various fonts and the extracted text preserves abbreviations (*ẽ*…) and special characters (*ſ*…) but not ligatures (*fi*…).

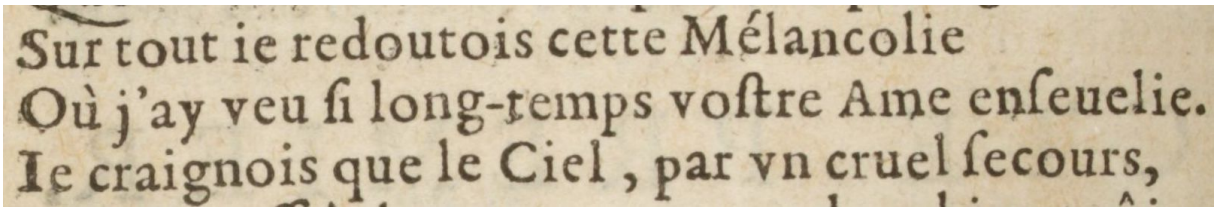
Sur tout ie redoutois cette Mélancolie
Où j'ay veu ſi long-temps voſtre Ame enſeuelie.
Ie craignois que le Ciel, par vn cruel ſecours,

Fig. 1 Racine, *Andromaque*, Paris, BNF, RES-YF-3206, p. 2

Pre-processing

Following previous successful experiments (Bollmann, 2012), a rule-based system for pre-orthographic French has been developed (Riguet, 2019). It is based on two lexical resources: Morphalou, an open lexical database of inflected forms of contemporary French (Romary et al., 2004), and LGeRM, an open morphological lexicon for middle French (Souvay and Pierrel, 2009) now covering also 17th c. French (Diwersy et al., 2017). Based on these two databases, the normaliser applies transformations on each token, before a manual correction of the result.

Normalisation consists of aligning 17th c. graphic systems (source) to 21st c. orthography (target)

Source	Target
Sur tout ie redout**o**is cette Mélancolie	**Surtout j**e redout**a**is cette Mélancolie
Où j'a**y** v**eu f**i long**-**temps v**of**tre **A**me ense**u**elie.	Où j'a**i** v**u s**i longtemps v**o**tre **Â**me ensevelie.
Ie craign**o**is que le Ciel, par **v**n cruel **f**ecours,	**J**e craign**a**is que le Ciel, par **u**n cruel **s**ecours,

First results with an NMT-based normaliser

We have decided to use NMTPYTORCH (Caglayan et al., 2017). The baseline model is composed of a 2-layer bi-directional GRU [14] encoder and a 2-layer conditional GRU (Cho et al., 2014) decoder with MLP attention (Cho et al., 2014). The encoder and the decoder both have 256 hidden units and their initial hidden state is initialised to 0. The embedding dimensionality is also set to 256. Two versions of the system have been trained. The first one is a word level system and the second one uses the byte pair encoding (BPE) (Sennrich et al., 2015) which operates at the subword level. The corpus has been divided into two parts: 90% of the lines have been used for training and 10% for testing.

	Lines	Tokens	Characters
Train	3,133	5,6825	348,098
Test	346	5,959	37,000
Total	3,479	62,784	385,098

Five trainings have been made with different initialisations on the two different models: words and subwords (*i.e.* BPE units). Accuracy of the result is calculated with BLEU scores (Papineni et al., 2002).

Model	Average BLEU	Best BLEU
Words	79.27	82.960
BPE	75.79	77.070

These BLEU scores still have to be used with extreme care considering the limited size of our corpus. They are however promising enough to engage in the production of a large-scale corpus for a NMT-based normaliser.

Future developments

To be as universal as possible, our training data must reflect all the lexical and graphic variety of 17th c. French. We are therefore engaging in the construction of a representative corpus of modern French, including excerpts of literary (plays, novel, poems…) and non-literary texts (theology, medicine, law, science…), in verse and in prose, spread diachronically across the century, and taken from original editions, reprints or illegal prints. Along this compilation phase, the OCR model and the rule-based normalising solution will be regularly improved to increase their efficiency before a final open source release.

The final corpus, expanded with back translation (Domingo and Casacuberta, 2018b), will be used for the training of an NMT-based solution. On top of words and subwords, character-level NMT will also be tested to provide the most efficient tool. A special model, trained to normalise the result of the rule-based system rather than the raw OCRised text will be tested, to test the efficiency of a hybrid system using both technologies.

Bibliography

Baron, A. and Rayson, P. (2008). VARD2 : a Tool for Dealing with Spelling Variation in Historical Corpora. *Postgraduate Conference in Corpus Linguistics*. Birmingham, UK http://eprints.lancs.ac.uk/41666/ (accessed 3 December 2018).

Bollmann, M. (2012). (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*. Lisbon, Portugal, pp. 3–14 https://www.linguistics.ruhr-uni-bochum.de/comphist/pub/acrh12.pdf.

Bollmann, M. and Søgaard, A. (2016). Improving Historical Spelling Normalization With Bi-Directional LSTMs and Multi-Task Learning. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, pp. 131–39 http://arxiv.org/abs/1610.07844.

Caglayan, O., García-Martínez, M., Bardet, A., Aransa, W., Bougares, F. and Barrault, L. (2017). NMTPY: A Flexible Toolkit for Advanced Neural Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*, **109**(1): 15–28 doi:10.1515/pralin-2017-0035.

Cho, K., Merrienboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 1724–34 http://arxiv.org/abs/1406.1078.

Diwersy, S., Falaise, A., Lay, M.-H. and Souvay, G. (2017). Ressources et méthodes pour l'analyse diachronique. *Langages*, **N° 206**(2): 21–44.

Domingo, M. and Casacuberta, F. (2018a). Spelling Normalization of Historical Documents by Using a Machine Translation Approach. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*. Alicante, Spain, pp. 129–37 http://rua.ua.es/dspace/handle/10045/76035.

Domingo, M. and Casacuberta, F. (2018b). A Machine Translation Approach for Modernizing Historical Documents Using Back Translation. *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*. Bruges, Belgium, pp. 39–47 https://workshop2018.iwslt.org/downloads/Proceedings_IWSLT_2018.pdf.

Domingo, M., Chinea-Rios, M. and Casacuberta, F. (2017). Historical Documents Modernization. *The Prague Bulletin of Mathematical Linguistics*, **108**: 295–306.

Gabay, S. (2019). OCRising 17th French prints. *E-Ditiones* https://editiones.hypotheses.org/1958.

Guez de Balzac, J.-L. (1624). *Lettres Du Sieur de Balzac*. Paris: T. Du Bray https://catalogue.bnf.fr/ark:/12148/cb300515241.

Korchagina, N. (2017). Normalizing Medieval German Texts: from rules to deep learning. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Gothenburg, Sweden, pp. 12–17 http://aclweb.org/anthology/W17-0504.

Ljubešic, N., Zupan, K., Fišer, D. and Erjavec, T. (2016). Normalising Slovene data: historical texts vs. user-generated content. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. Bochum, Germany, pp. 146–155 https://www.linguistics.rub.de/konvens16/pub/19_konvensproc.pdf.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, USA, pp. 311–318 doi:10.3115/1073083.1073135.

Porta, J., Sancho, J.-L. and Gomez, J. (2013). Edit Transducers for Spelling Variation in Old Spanish. *Proceedings of the Workshop on Computational Historical Linguistics (NoDaLiDa 2013)*. Oslo, Norway, pp. 70–79 http://www.ep.liu.se/ecp/087/006/ecp1387006.pdf.

Racine, J. (1668). *Andromaque*. Paris: Barbin https://catalogue.bnf.fr/ark:/12148/cb38651697n.

Riguet, M. (2019). *Normalisa, Script à Base de Règles Pour Normaliser Les Textes Français Du XVIe Au XIXe Siècle*. https://mriguet.github.io/Normalisa (accessed 28 April 2019).

Romary, L., Salmon-Alt, S. and Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. *The 20th International Conference on Computational Linguistics (COLING 2004) - ElectricDict '04 Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*. Geneva, Switzerland, pp. 22–28 https://hal.inria.fr/inria-00121489.

Scherrer, Y. and Erjavec, T. (2013). Modernizing Historical Slovene Words with Character-Based SMT. *4th Biennial Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*. Sofia, Bulgaria, pp. 58–62 https://hal.inria.fr/hal-00838575 (accessed 3 December 2018).

Sennrich, R., Haddow, B. and Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pp. 1715–1725 https://arxiv.org/abs/1508.07909 (accessed 3 December 2018).

Souvay, G. and Pierrel, J.-M. (2009). LGeRM Lemmatisation des mots en Moyen Français. *Traitement Automatique Des Langues*, **50**(2): 149–72.

Tjong Kim Sang, E., Bollman, M., Boschker, R., Casacuberta, F., Dietz, F. M., Dipper, S., Domingo, M., et al. (2017). The CLIN27 Shared Task : Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation. *Computational Linguistics in The Netherlands Journal*, **7**: 53–64.